Contributed Articles

Optimal Difference Fourier Synthesis in Fibre Diffraction

R.P.Millane and S.Baskaran

Whistler Center for Carbohydrate Research and Computational Science and Engineering Program Purdue University West Lafayette, Indiana 47907-1160 U.S.A.

A weighting scheme for difference Fourier synthesis in fibre diffraction that yields a minimum mean-square-error estimate for the missing electron density is described. Simulations show the advantages of using the weighting scheme, and other applications in fibre diffraction are discussed.

1. Introduction

Difference Fourier synthesis is an important technique in small molecule and protein crystallography, where it is used to locate missing components (*e.g.* counterions, or solvent or other bound molecules) or to correct errors (such as in side-chain positions), in crystal structures. Since difference Fourier synthesis involves using a partially determined structure to phase the diffraction data, it is also closely related to molecular replacement methods. For single crystal crystallography, difference Fourier methods are quite well established. The most straightforward approach is to synthesise a map of the difference electron density by inverse Fourier transforming the difference between the observed (measured) amplitudes $|F_{\mathbf{h}}^{o}|$ and the calculated amplitudes (*i.e.* those calculated $|F_{\mathbf{h}}^{c}|$ from the known part of the structure, or from a model structure), phased by the known part, *i.e.* $\Delta\rho(x) = \mathcal{F}^{-1}\{(|F_{\mathbf{h}}^{o}|-|F_{\mathbf{h}}^{c}|)\exp(i\phi_{\mathbf{h}}^{c})\}$, where $\mathcal{F}\{\cdot\}$ denotes the Fourier transform and $\phi_{\mathbf{h}}^{c}$ denotes the phase of $F_{\mathbf{h}}^{c}$. This difference map can be improved upon by weighting the $|F_{\mathbf{h}}^{o}|$ in accordance with the reliability of the $\phi_{\mathbf{h}}^{c}$, and by including the effects of errors in the partial structure [1,2]. Difference Fourier methods have also found significant application in fibre diffraction analysis since the early days for analysis of the symmetry of nucleic acids [3], through detailed visualization of solvent and cation interactions in polysaccharide systems [4] and applications to helical viruses [5]. However, the relatively straightforward approach to difference synthesis in single crystal crystallography does not translate exactly to fibre diffraction analysis, and applications have involved approximations unless special symmetry was present. The difficulty arises because one needs to synthesise a three-dimensional electron density function but, whereas the model structure provides the quantities $|F_{\mathbf{h}}^{c}|$ and $\phi_{\mathbf{h}}^{c}$, the two-dimensional diffraction data available from a fibre diffraction experiment do not generally provide the individual quantities $|F_{\mathbf{h}}^{o}|$. To calculate a fibre diffraction difference map, one must use the form

$$\Delta\rho(\mathbf{x}) = \mathcal{F}^{-1}\left\{\left(\left|F_{\mathbf{h}}^{o'}\right|-\left|F_{\mathbf{h}}^{c}\right|\right)\exp\left(i\phi_{\mathbf{h}}^{c}\right)\right\} \qquad (1)$$

where the $|F_{\mathbf{h}}^{o'}|$ are approximations to the $|F_{\mathbf{h}}^{o}|$ that must be estimated from the fibre diffraction data.

Two approaches to this problem have traditionally been used in fibre diffraction analysis. For a polycrystalline specimen, the observed intensity of a diffraction spot I^{o} takes the form

$$I^{o} = \sum_{\mathbf{h}}\left|F_{\mathbf{h}}^{o}\right|^{2} \qquad (2)$$

where the sum is over the reflections $\mathbf{h}$ that overlap in the spot, as a result of either systematic or accidental overlap. One approach to approximating $|F_{\mathbf{h}}^{o}|$ is to assume that all the contributing reflections are of equal amplitude so that

$$\left|F_{\mathbf{h}}^{o'}\right| = \sqrt{\frac{I^{o}}{\mathcal{N}}} \qquad (3)$$

where $\mathcal{N}$ is the number of reflections contributing to I^{o}. This $|F_{\mathbf{h}}^{o'}|$ is used in equation (1). This is a parsimonious choice as it uses phase information from the model (or known part of the structure), but does not make use of the relative values of the amplitudes $|F_{\mathbf{h}}^{c}|$ from the model in estimating $|F_{\mathbf{h}}^{o'}|$. This is the approach usually taken as it reduces an overriding concern, that of biasing the result towards the model. The second approach is to make use of the structure amplitudes $|F_{\mathbf{h}}^{c}|$ derived from the model to estimate $|F_{\mathbf{h}}^{o'}|$ and assume that the observed amplitudes are distributed in the same ratios, but are constrained such that they satisfy equation (2), *i.e.*

$$\left|F_{\mathbf{h}}^{o'}\right| = \frac{\left|F_{\mathbf{h}}^{c}\right|}{\sqrt{I^c}}\sqrt{I^o} \qquad (4)$$

where I^c represents the fibre diffraction intensities that would be observed from the partial structure, *i.e.* $I^c = \sum_{\mathbf{h}} |F_{\mathbf{h}}^c|^2$. This approach makes use of more information from the model (*i.e.* the amplitude ratios as well as the phases) and is therefore expected to be more reliable if the model is closer to the full structure (*i.e.* $\phi_{\mathbf{h}}^c$ is closer to the phase of the full structure and $|F_{\mathbf{h}}^c|$ is closer to $|F_{\mathbf{h}}^o|$). However, if the model is not close to the full structure, this synthesis can introduce more bias towards the model structure than does the first approach described above. Therefore, one is generally cautious about using this synthesis unless one is confident that the partial structure is a good approximation to the full structure, *i.e.* the missing part constitutes a relatively small part of the full structure.

Returning to single crystal crystallography, in 1960 Sim [1] showed that an improved estimate for the missing structure can be obtained by using

$$\Delta\rho(\mathbf{x}) = \mathcal{F}^{-1}\left\{\left(w\left|F_{\mathbf{h}}^{o}\right| - \left|F_{\mathbf{h}}^{c}\right|\right)\exp\left(i\phi_{\mathbf{h}}^{c}\right)\right\} \qquad (5)$$

where the weighting function w is given by $w = I_1(2|F_{\mathbf{h}}^o||F_{\mathbf{h}}^c|/\Sigma)/I_0(2|F_{\mathbf{h}}^o||F_{\mathbf{h}}^c|/\Sigma)$ for the acentric reflections and $w = \tanh(|F_{\mathbf{h}}^o||F_{\mathbf{h}}^c|/\Sigma)$ for the centric reflections. $I_m(\cdot)$ is the modified Bessel function of the first kind, $\Sigma = \sum_j f_j^2$ and the f_j are the scattering factors of the missing atoms in the unit cell. Since the missing atoms are generally not known, Σ must be estimated from the data. The Sim weighting function stems from minimizing the difference (in the least-squares sense) between the actual and estimated missing electron density, and is based on Wilson's statistics [6] for the distribution of the structure amplitudes for the missing electron density. Applying Sim weighting can give a significant improvement in the accuracy of a difference map. Further modifications can be made to the synthesis to take into account errors in the known part of the structure [2].

The amplitude of a difference map is scaled by approximately one-half (for the acentric reflections), so that a map of the full structure based on the coefficients $(2w|F_{\mathbf{h}}^o|-|F_{\mathbf{h}}^c|)\exp(i\phi_{\mathbf{h}}^c)$ gives better relative peak heights than does one based on $w|F_{\mathbf{h}}^o|\exp(i\phi_{\mathbf{h}}^c)$ [7]. Note, however, that for the centric reflections the coefficients $w|F_{\mathbf{h}}^o|\exp(i\phi_{\mathbf{h}}^c)$ should be used. Namba and Stubbs [8] have shown that a (unweighted) synthesis of the full electron density based on the coefficients $(m|F_{\mathbf{h}}^{o'}|-(m-1)|F_{\mathbf{h}}^c|)\exp(i\phi_{\mathbf{h}}^c)$, where m is the number of degrees of freedom in the spot to which $|F_{\mathbf{h}}^o|$ belongs (defined in the next section), gives the correct relative peak heights in the fibre diffraction case.

Optimal fibre diffraction difference Fourier synthesis

We have carried out an analysis [9,10] analogous to that of Sim for the fibre diffraction case using the statistics for the fibre diffraction data I^o [11,12] (analogous to Wilson's statistics). Significant insight is obtained by formally posing the problem as a Bayesian estimation problem [9]. The synthesis that minimizes the mean-square-error (*i.e.* the MMSE estimate) to the actual electron density is derived as follows. The structure factors corresponding to the MMSE estimate of the missing electron density are given by using the *posterior mean* $\langle F^o\, P(F^o|I,F^c)\rangle$ for $F^{o'}$ in equation (1), where $P(F^o|I,F^c)$ is the posterior (conditional) probability density for the observed structure factors given the intensity data and the structure factors of the partial structure. Bayes theorem is used to derive the posterior density from $P(I|F^o,F^c)$ and a prior density for the structure factors of the missing part of the structure based on Wilson's statistics, and using equation (2). Evaluating the above expressions then shows that the MMSE is given by using [9,10]

$$F_{\mathbf{h}}^{o'} = w_m \frac{\left|F_{\mathbf{h}}^{c}\right|}{\sqrt{I^c}}\sqrt{I^o} \qquad (6)$$

in equation (1), where

$$w_m = \frac{I_{m/2}\left(c\sqrt{I^o I^c}/\Sigma\right)}{I_{m/2-1}\left(c\sqrt{I^o I^c}/\Sigma\right)} \qquad (7)$$

m is the number of degrees of freedom in the datum I^o (twice the number of contributing acentric reflections or the number of contributing centric reflections), $c = 1$ for centric reflections and 2 for acentric reflections, and Σ is defined as in the previous section. We have devised effective methods for estimating Σ from fibre diffraction data [10]. We note that equation (7) is strictly correct only if the datum I^o contains only reflections of one type (either centric or acentric). The weighting function is considerably more complicated for mixed data [10]. Note that equation (6) takes the form of equation (4) except for the weight, *i.e.* the amplitudes are divided in the same proportion as the amplitudes from the known part, but are weighted down more, the more the known part deviates from the full structure (*i.e.* the more dissimilar are I^o and I^c). The weights depend on the number of degrees of freedom of each datum, and reduce (as they must) to the Sim weights for $m = 1,2$. The weighting function, $w_m(\chi)$, as a function of $\chi = c\sqrt{(I^oI^c)}/\Sigma$, for different numbers of degrees of freedom is shown in Fig. 1. As expected, w_m decreases as the derived $F_{\mathbf{h}}^{o'}$ become less reliable, *i.e.* with decreasing χ and increasing m. The effects of errors in the data and in the known part of the structure can also be included in the analysis [10].

Furthermore, we have also shown that the two currently used syntheses described above correspond to *maximum a posterior* (MAP) estimates, *i.e.* estimates that are located at the maxima of certain posterior densities [9]. The estimate equation (3) corresponds to is the value of $F_{\mathbf{h}}^{o'}$ that maximizes $P(F_{\mathbf{h}}^o|I,\phi_{\mathbf{h}}^c)$, *i.e.* it uses the intensity data, and the phases (but not the amplitudes) derived from the partial structure. The second estimate, equation (4), utilizes both the amplitude and phase from the partial structure, and corresponds to the maximum of $P(F_{\mathbf{h}}^o|I,F_{\mathbf{h}}^c)$. The MMSE estimate is generally better than the MAP estimates since it is based on the mean rather than the mode.

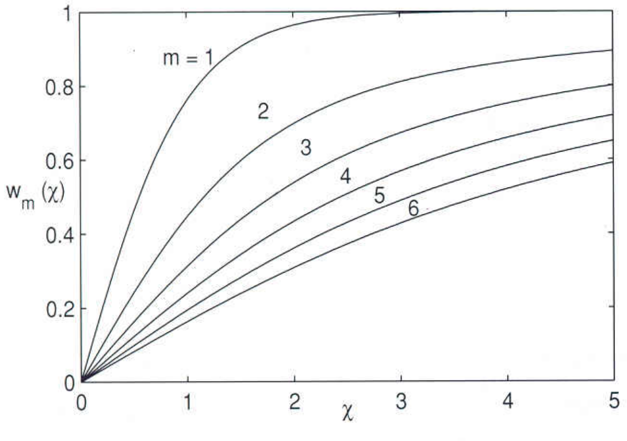

Figure 1: The weighting function $w_m(\chi)$ as a function of χ=c$\sqrt{(I^oI^c)}/\Sigma$, for different values of m.

Simulations

The performance of the various difference Fourier syntheses described above was investigated by calculating difference maps using simulated fibre diffraction data for mannan II, the structure of which is described by Millane and Hendrixson [13]. The unit cell is orthorhombic and the space group is $I222$. Four polymer chains pass through the unit cell with their axes at (1/4, 1/4), (1/4, 3/4), (3/4, 1/4) and (3/4, 3/4) in the *a-b* plane. In determining this structure, conventional difference Fourier synthesis was used to locate an ordered water molecule in the crystal structure [13].

Synthetic fibre diffraction data I^o to 2Å resolution were calculated based on the crystal structure consisting of the polymer and water molecules (represented by oxygen atoms). The partial structure was taken to be that consisting of the polymer molecules only, from which the structure factors $F_{\mathbf{h}}^c$ were calculated. The water molecule accounts for approximately 10% of the electrons in the unit cell. There are 71 unique reflections within this resolution limit, that give 51 fibre diffraction data as a result of both systematic and accidental overlaps. There are no mixed data, *i.e.* all data contain either centric or acentric reflections. For these simulations, Σ was calculated explicitly, as a function of resolution, using the atomic scattering factors of the missing atoms. Using these data, difference electron density maps for the water molecules were calculated using the three different methods described above.

Contour maps of the true difference electron density (calculated using the $F_{\mathbf{h}}^o$) are shown in Fig. 2a at the $z = 0$ level (left) where two water molecules are located as shown by two strong peaks, and at the $z = 1/4$ level (right), which is between the water molecules and the difference map is relatively featureless (below the lowest contour level). Difference maps calculated based on equations (6) (MMSE), (4) and (5) are shown in Figs. 2b, c and d, respectively. Comparison of these maps with the true map (Fig. 2a) shows that the MMSE map (Fig. 2b) is superior to the other two maps in terms of a higher amplitude of the peaks corresponding to the water molecules, lower amplitudes of spurious peaks, and a smaller overall noise level. The improvement over the maps obtained using the conventional fibre

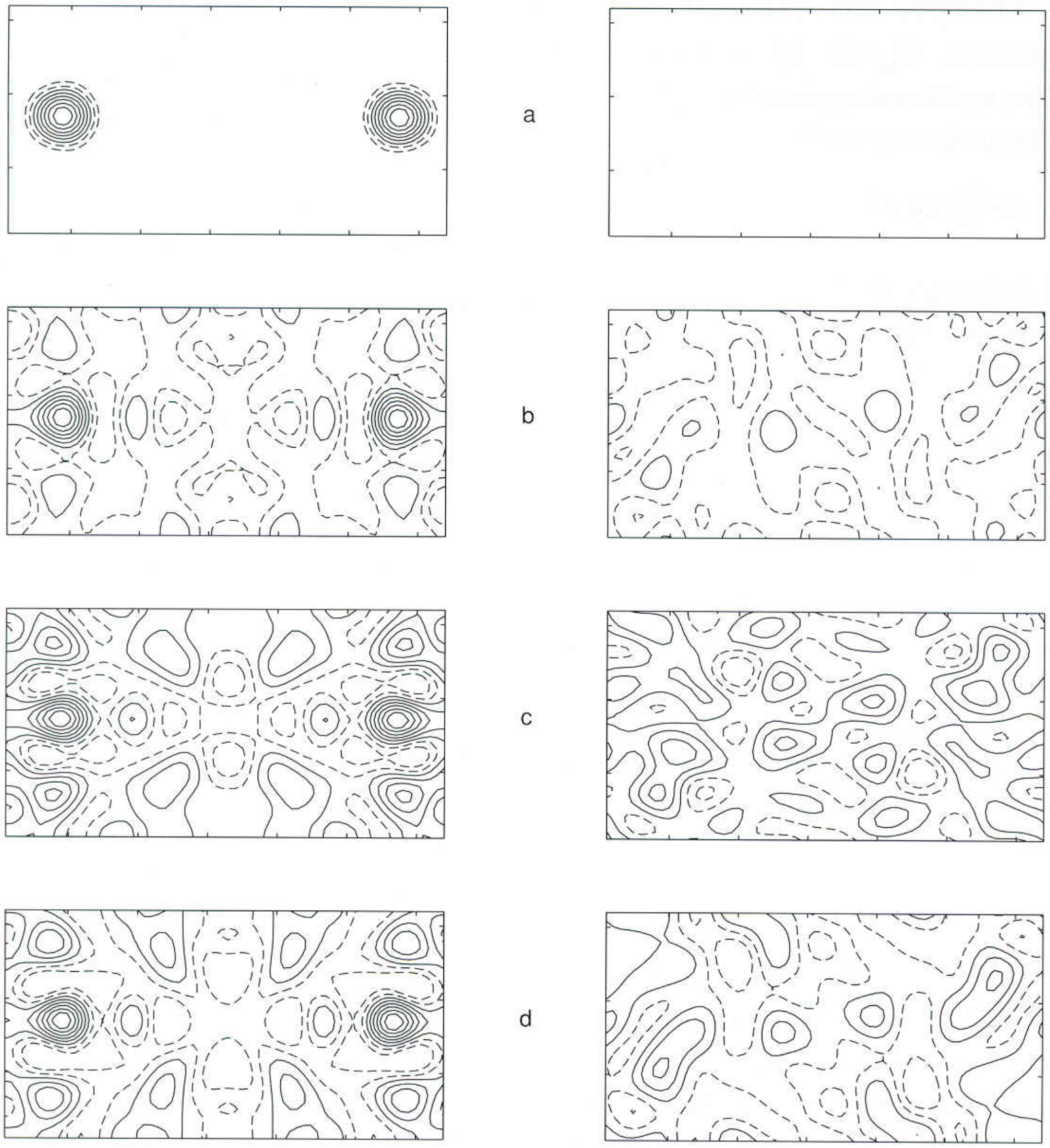

Figure 2: Contour plots of (a) the true difference electron density map, and estimated maps using (b) equation (6), (c) equation (3), and (d) equation (4), for mannan II. The maps are shown at levels $z=0$ (left) and $z=1/4$ (right). Dashed contours denote negative levels.

diffraction difference syntheses (Figs. 2c and d) is significant and is expected to improve map interpretability when real data are used and the missing structure is more complex than in this example. Note that in this example, the synthesis using equation (4) (Fig. 2d) is superior to that obtained using equation (3) (Fig. 2c). This is probably because the fraction of the total structure that is missing is rather small so that the ratios of the $|F_{\mathbf{h}}^{c}|$ within a spot is a good approximation to the ratios of the $|F_{\mathbf{h}}^{o}|$.

Discussion

A rigorous weighting scheme for difference Fourier synthesis in fibre diffraction has been derived, and is a generalisation of Sim's weighting scheme. Simulations show the potential for significant improvements in the interpretability of difference maps using this scheme. Extensions to include the effects of errors in the diffraction data and in the partial structure, as well as the treatment of mixed (centric and acentric) data, are in progress. The weighting scheme, as well as traditional unweighted syntheses in fibre diffraction, can be neatly categorised in a Bayesian framework.

The analysis described above applies, strictly, to diffraction data from polycrystalline fibres. However, a very similar approach should be applicable to continuous diffraction data from non-crystalline fibres also. In the latter case, the problem is one of estimating the difference electron density from the continuous diffraction data $I_l^o(R)$ and the Fourier-Bessel structures $G_{nl}^c(R)$ of a known partial structure. The traditional approach is to estimate the missing electron density as [8]

$$\Delta\rho(r,\theta,z) = \mathcal{FB}^{-1}\left\{\left(\left|G_{nl}^{o'}(R)\right| - \left|G_{nl}^{c}(R)\right|\right)\exp\left(i\phi_{nl}^{c}(R)\right)\right\} \quad (8)$$

where $\mathcal{FB}\{\cdot\}$ denotes the Fourier-Bessel transform, $|G_{nl}^{o'}(R)|$ is an approximation to $|G_{nl}^o(R)|$, and $\phi_{nl}^c(R)$ is the phase of $G_{nl}^c(R)$. The diffracted intensity $I_l(R)$ is given by

$$I_l(R) = \sum_n \left|G_{nl}^o(R)\right|^2 \quad (9)$$

where the sum is over the values of n that satisfy the helix selection rule. In principle, the number of terms in the sum is infinite. However, for a particular molecule and value of R, the $G_{nl}(R)$ are small for n larger than a fixed value, and the number of terms contributing to equation (9) is effectively finite [14]. Equation (9) is then of the same form as equation (2), and currently used methods to approximate $|G_{nl}^o(R)|$ are analogous to those described in section 1, *i.e.* dividing $I_l^o(R)$ up either equally among the contributing $|G_{nl}^o(R)|^2$, or in the same proportion as $|G_{nl}^c(R)|^2$ divides $I_l^c(R)$. The $G_{nl}(R)$ approximately follow Wilson's statistics [11] so that an identical approach to that described in section 2 leads to the same weighting scheme for continuous diffraction data. While this is likely to be effective in practice, a more rigorous study of the effects of a sharp cutoff in the order of the Fourier-Bessel terms that contribute to the right-hand-side of equation (9), and deviations from Wilson's statistics for the $|G_{nl}(R)|$, would be worthwhile.

Molecular replacement is becoming an increasingly important approach in fibre diffraction. This is particularly the case for large systems where isomorphous replacement is very demanding, but where structure determination by *ab initio* model building is not feasible. In such cases, the use of a related structure to phase, and separate the amplitudes of, the diffraction data, or to provide an initial model, is attractive. Since the model structure in molecular replacement can be considered a partial structure (with errors), molecular replacement can be considered as a problem of difference Fourier synthesis. In fact, current applications of molecular replacement in fibre diffraction (almost exclusively with continuous diffraction data, e.g. [15]) use the model structure to phase and separate the structure factors contributing to the diffraction data from the unknown structure, in the same way as for traditional difference Fourier synthesis as described above. The weighting scheme described here is therefore expected to be useful in molecular replacement also. Two considerations will be important. Since the model structure is not strictly a part of the full structure, the effects of errors in the partial structure will be more important. And since one wishes to synthesise the full unknown structure, the effects of bias towards the model will be more important. Applications of the results presented in section 2 to molecular replacement, and investigation of the above considerations, would be worthwhile.

Supported by the U.S. National Science Foundation (DBI-9722862)

References

[1] G.A.Sim. *Acta Cryst.*, **13**, 511-512, 1960.
[2] R.J.Read, *Acta Cryst.*, **A42**, 140-149, 1986.
[3] S.Arnott, M.H.F.Wilkins, W.Fuller, and R.Langridge, *J. Mol. Biol.*, **27**, 535-548, 1967.
[4] J.J.Cael, W.T.Winter, and S.Arnott, *J. Mol. Biol.*, **125**, 21-42, 1978.
[5] K.Namba, R.Pattanayek, and G.J.Stubbs, *J. Mol. Biol.*, **208**, 307-325, 1989.
[6] A.J.C.Wilson, *Acta Cryst.*, **2**, 318-321, 1949.
[7] P.Main, *Acta Cryst.*, **A35**, 779-785, 1979.
[8] K.Namba and G.J.Stubbs, *Acta Cryst.*, **A43**, 533-539, 1987.
[9] S.Baskaran and R.P. Millane, *Proc SPIE*, **3170**, 227-237, 1997.
[10] S.Baskaran and R.P.Millane, *Acta Cryst.*, in preparation, 1997.
[11] G.Stubbs, *Acta Cryst.*, **A45**, 254-258, 1989.
[12] R.P.Millane, *Acta Cryst.*, **A46**, 552-559, 1990.
[13] R.P.Millane and T.L.Hendrixson, *Carbohdr. Polym.*, **25**, 245-251, 1994.
[14] L.Makowski, *J. Appl. Cryst.*, **15**, 546-557, 1982.
[15] S.Lobert and G.Stubbs, *Acta Cryst.*, **A46**, 993-997, 1990.